Nucleic Acids Research, 2004, Vol. 32, Database issue ***D315–D318***
DOI: 10.1093/nar/gkh091

Génolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts

David Sherman*, Pascal Durrens[1], Emmanuelle Beyne, Macha Nikolski and Jean-Luc Souciet[2] for the Génolevures Consortium

LaBRI, Laboratoire Bordelais de Recherche en Informatique, UMR CNRS 5800, 351 cours de la Libération, 33405 Talence cedex, France, [1]Centre de Bioinformatique de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux, France and [2]Génétique et Microbiologie, FRE 2326 ULP/CNRS, GDR CNRS 2354, Institut de Botanique, 28 rue Goëthe, 67000 Strasbourg, France

Received August 21, 2003; Revised and Accepted October 7, 2003

ABSTRACT

The Génolevures online database (http://cbi.labri.fr/Genolevures/) provides data and tools to facilitate comparative genomic studies on hemiascomycetous yeasts. Now, four complete genome sequences recently determined (*Candida glabrata*, *Kluyveromyces lactis*, *Debaryomyces hansenii*, *Yarrowia lipolytica*) have been added to the partial sequences of 13 species previously analysed by a random approach. The database also includes the reference genome *Saccharomyces cerevisiae*. Data are presented with a focus on relations between genes and genomes: conservation of genes and gene families, speciation, chromosomal reorganization and synteny. The Génolevures site includes a community area for specific studies by members of the international community.

INTRODUCTION

With their relatively small and compact genomes, yeasts offer a unique opportunity to explore eukaryotic genome evolution by comparative analysis of several species. Yeasts are widely used as cell factories, for the production of beer, wine and bread and more recently of various metabolic products such as vitamins, ethanol, citric acid, lipids, etc. Yeasts can assimilate hydrocarbons (genera *Candida*, *Yarrowia* and *Debaryomyces*), depolymerize tannin extracts (*Zygosaccharomyces rouxii*), and produce hormones and vaccines in industrial quantities through heterologous gene expression. For review see (1). Several yeast species are pathogenic for humans. Among the most frequent disease agents are the Hemiascomycetes *Candida albicans*, *Candida glabrata*, *Candida tropicalis* and the Basidiomycete *Cryptococcus neoformans*. Even *Saccharomyces cerevisiae* may be pathogenic in immunocompromised patients. The most well known yeast in the Hemiascomycete class is *S.cerevisiae,* widely used as a model organism for molecular genetics and cell biology studies, and as a cell factory. As the most thoroughly annotated genome of the small eukaryotes, it is a common reference for the annotation of other species. The hemiascomycetous yeasts represent a homogeneous phylogenetic group of eukaryotes with a relatively large diversity at the physiological and ecological levels. Comparative genomic studies within this group have proved very informative (2–4).

The Génolevures programme is devoted to large-scale comparisons of yeast genomes from various branches of the Hemiascomycete class, with the aim of addressing basic questions of molecular evolution such as the degree of gene conservation, the identification of species-specific, clade-specific or class-specific genes, the distribution of genes among functional families, the rate of sequence and map divergences, and mechanisms of chromosome shuffling.

The focus of the Génolevures database is the relations between genes and genomes. We curate relations of orthology and paralogy between genes, as individuals or as members of gene families, relations of synteny and duplication between chromosomal segments, and gain and loss of genes and functions. We do not provide detailed annotations of individual genes and proteins of *S.cerevisiae*, which are already carefully maintained by MIPS (http://mips.gsf.de/projects/fungi/), CYGD (5) and SGD (http://www.yeastgenome.org/) (6) as well as in general-purpose databases such as Swiss-Prot and EMBL.

The Génolevures website provides data, text and graphic search tools, and community pages for ongoing development. Since its inception in December 2000, the site has been developed by a mixed team of biologists and computer scientists working in close collaboration. The various changes that the site has undergone have all been motivated by extensive feedback provided by members of the Génolevures Consortium.

ORIGIN OF THE DATA

The core of the Génolevures database is a large set of novel DNA sequences. The Génolevures sequencing project was

*To whom correspondence should be addressed. Tel: +33 540 00 6922; Fax: +33 540 00 6669; Email: sherman@labri.fr

The Génolevures Consortium is coordinated by J. L. Souciet and is composed of laboratories from the Institut Pasteur (Paris), the INA-PG (Paris-Grignon), the Universities Bordeaux 1 and 2, Claude Bernard (Lyon), Paris-Sud (Orsay), Pierre et Marie Curie (Paris 6) and Louis Pasteur (Strasbourg), the Institut Curie (Paris), the Génoscope (Evry) and the Génopole Pasteur-Ile-de-France (Paris)

undertaken in two phases: first, the random sequencing of 13 species distributed throughout the Hemiascomycete class, then the complete sequencing of four selected species representative of major clades inside this class. The initial random sequencing exploration aimed specifically at providing a broader, but consequently shallower, view of evolution than could be provided by the complete genomes of just a few related species. A set of species representing the various branches of the Hemiascomycetes class was defined. Single-pass sequencing of both ends of short genomic fragments led to ~50 000 RST sequences (2500–5000 per species) that were first compared with *S.cerevisiae* rDNA, tRNA genes, Ty elements and mitochondrial sequences, and then with a non-redundant compilation of protein sequences from major data banks. Every alignment produced by BLAST (7) was manually evaluated and validated by experts and 46 600 homologies were recorded for the RSTs, which were annotated using *S.cerevisiae* data extracted from MIPS. Roughly 20 000 new yeast genes were thus found for the 13 species, and the *S.cerevisiae* genome was revised by the addition of 50 novel genes and 26 gene extensions.

On the basis of these results and other considerations, four selected species have now been systematically sequenced and annotated: *Yarrowia lipolytica*, which assimilates hydrocarbons and produces citric acid from *n*-alkanes, vegetable oils or glucose under aerobic conditions, *Debaryomyces hansenii*, which is a halotolerant yeast that also assimilates hydrocarbons, *Kluyveromyces lactis*, which overproduces and secretes the aspartyl protease chymosin, the active constituent of the cheese rennet, and *C.glabrata*, the second most common human pathogen causing candidiasis. Roughly 29 000 protein coding genes have been identified. The systematic *de novo* annotation of these genes and other genetic elements is an ongoing process.

All DNA data made available through the Génolevures website are public and are deposited in the EMBL data bank.

In silico analyses play an essential role for understanding the relations between yeast genomes. Different methods are applied: (i) gene families within species are identified using the partitioning and clustering method described in (8); (ii) orthologous families between species are identified with the MCL Markov clustering method (9,10); (iii) synteny conservation and chromosomal maps are analysed using the method of (11) and the ADHoRE method (12).

Data from external sources are integrated in the Génolevures database when appropriate but detailed annotations of elements curated elsewhere are only maintained as database cross-references that 'link out' from our web pages to external sites. For *S.cerevisiae*, functional annotations are taken from MIPS, and gene synonyms from SGD.

COMMUNITY

Génolevures is an active community of yeast researchers based on several French laboratories and an informal network of international partners. This open community contributes both through participation in genome annotation, and through the communication of specific results. The Génolevures website offers to the yeast community an area for presentations of specific studies, where supplementary data from published results can be disseminated and explored through searchable indexes and other pertinent links.

DATABASE ORGANIZATION

The Génolevures database is organized using a relational model with a set of carefully designed extensible ontologies. We chose to base the model on extensible ontologies rather than a fixed object-oriented model as it is more flexible, despite the risk of reification errors. The principal ontology for Génolevures has three main branches. The branch 'sequence-feature' uses a standard approach to modelling DNA and protein sequence features, and the branch 'evidence' describes the origin of the observation. The more novel 'relation' branch defines a vocabulary for describing relations between genes and groups of genes. Included are gene relations (e.g. homologies, family memberships), sequence relations (e.g. alignments, gene products), regulation relations (e.g. activation, repression) and interactions (e.g. binding). A new systematic gene nomenclature has been developed. It includes all possible genetic elements (protein-coding genes, RNA-coding genes, reiterated or degenerated sequences etc.) and allows for addition of newly annotated ones without altering the incremental numbering from left to right of chromosomes. The nomenclature also differenciates sequences from different strains or variants of the same species.

QUERY TOOLS AND DATA ACCESS

The Génolevures website (http://cbi.labri.fr/Genolevures/) provides tools for cross-species genome comparisons, access to alignments and annotation information, and access to complete sequence and annotation data produced by the project. These data may be consulted online, or downloaded in tabular or XML form. Queries provide easy answers to the most frequently asked questions. Predefined queries are presented by theme on the 'full search' page (http://cbi.labri.fr/Genolevures/Genolevures.php), and are summarized below.

(i) Is a given gene conserved in the class of Hemiascomycetous yeasts? (paragraph 'Search by ORF') Starting from a known gene name, systematic name or chromosome number, the first standard request produces a comparative table of homologues classified by species. Each column of the table represents a target species. For each gene, the value in that row for each column is either empty, indicating the absence of evidence of a homologue in that species, or a numerical value indicating the highest percentage identity of a validated alignment. In the case of a multigene family, a star is placed next to the value to indicate that it is the best of several values. Thus, at a glance, it is possible to see how the genes of interest are conserved across all hemiascomycetous species curated by Génolevures. Detailed homology information is provided by clicking on identity values.

(ii) To what extent is a given function represented in the class of hemiascomycetous yeasts? ('Search by annotation') Starting from a keyword or GO term (13) in the functional annotation of yeast ORFs, a standard request produces a list of sequences having that annotation. From this list it is possible to explore homologies, annotations, alignment results, and mapping to the *S.cerevisiae* chromosomes.

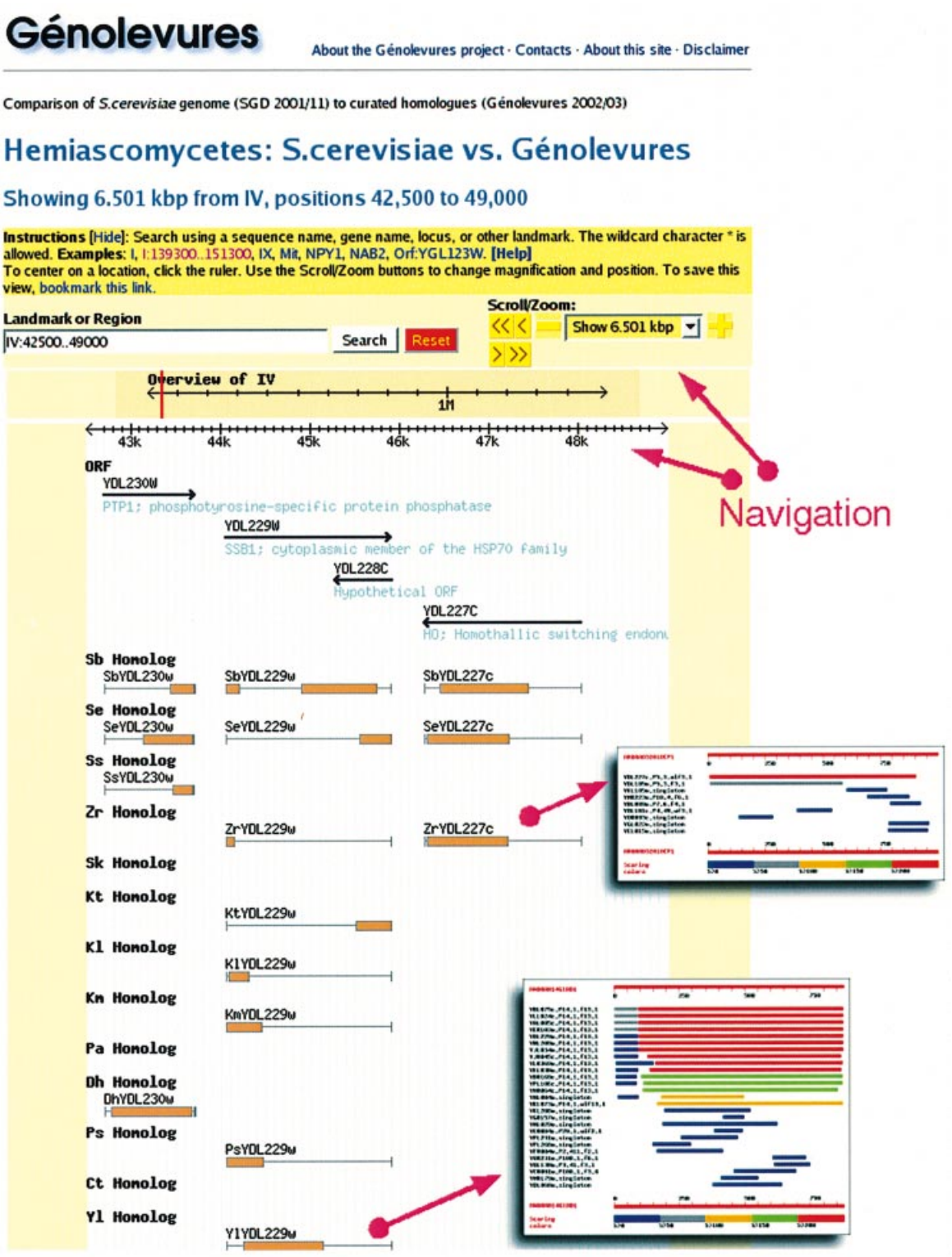

Figure 1. The neighborhood of *S.cerevisiae* YDL229w (SSB1) showing evidence of conservation of the gene in the hemiascomycetous yeasts. *YDL227c* (*HO*) seems not to be conserved beyond the genus *Kluyveromyces*. Clicking on an annotation in the image links out to evidence pages, including Blast alignments.

(iii) What detailed results are known for a given RST sequence? ('Search by RST') Full annotation data and Blast reports can be retrieved for a given sequence or clone identifier.

(iv) What is the level of synteny conservation in a defined chromosomal area? ('Search by RST,' button Map) A set of genes can also be mapped to an image showing the location on the *S.cerevisiae* genome of the corresponding homologues. In

this way one can visualize ruptures of synteny, or co-localization of groups of genes from another species.

(v) How do I obtain sequence data? ('Retrieve data') Nucleic acid sequences are available for published data, in Fasta, EMBL and Génolevures XML formats. All published nucleic acid sequences are also available in the EMBL data bank. Draft protein sequences for genomes still being annotated are also made available upon request.

Finally, the Génolevures website also provides a BLAST service to which one can submit DNA or protein sequences in FASTA format, and obtain alignment results against all Génolevures data, against selected species or against *S.cerevisiae*. Hyperlinks on the result page link back into the database.

GENOME BROWSER

For a given complete genome, it is possible to compare its localized sequence features with annotations induced from the relations of those features to other sequences. Using the Generic Genome Browser (CGB) (14), the Génolevures GGB service generates a clickable image that can be used to explore the query genome and its relations. Any of the Génolevures genomes can be taken as a reference. Annotation tracks show relations to other genomes. For example, Figure 1 represents the neighbourhood of the *S.cerevisiae YDL229w* (*SSB1*) gene, showing evidence that it is conserved throughout the Hemiascomycetes class. Hyperlinked data further suggest that it is a member of a multigenic family. *YDL227c* (*HO*), on the other hand, is not conserved in the genus *Kluyveromyces* or beyond.

ONGOING DEVELOPMENTS

Ongoing developments to the Génolevures database are currently concentrated on exploiting the data from the four new genome sequences produced by the consortium, on extending existing ontologies and on providing support for the distributed annotation effort. New graphical representations of genome relations and rearrangements are also being developed. Future developments under consideration will include the integration of other Hemiascomycete species sequenced by others.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We wish to thank all our colleagues from the Génolevures Consortium for numerous, friendly and creative discussions and for their devoted contributions to the sequencing, assembly and annotations of the yeast genomes. Special acknowledgement should be made to Lionel Frangeul for his installation and operation of CAAT-BOX package, to Ingrid Lafontaine and Emmanuel Talla for the analyses of gene families, Gilles Fischer for analysis of synteny conservation, and to Jean Weissenbach, Patrick Wincker and Christiane Bouchier of the Génoscope without whom the sequencing of yeasts would not have been possible. Hardware and technical support for Génolevures is provided by the Laboratoire Bordelais de Recherche en Informatique (LaBRI UMR 5800) for the Bordeaux Center for Bioinformatics, and is made possible by funding from the Aquitaine Région and the University Bordeaux 1. Génolevures is supported by CNRS (GDR 2354), by various sources from host institutions of participating laboratories and by CNRG through Génoscope and the réseau National des Génopoles.

REFERENCES

1. Kurtzmann,C.P. and Fell,J.W. (eds) (1998) *The Yeasts: A Taxonopmic Study*. 4th edn. Elsevier, Amsterdam.
2. Souciet,J. *et al*. (2000) Génolevures: Genomic exploration of the hemiascomycetous yeasts. Special Issue. *FEBS Lett.*, **487**, 1–149.
3. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
4. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
5. Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
6. Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M., Weng,S. and Botstein,D. (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.*, **26**, 73–79.
7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Tekaia,F., Blandin,G., Malpertuy,A., Llorente,B., Durrens,P., Toffano-Nioche,C., Ozier-Kalogeropoulos,O., Bon,E., Gaillardin,C., Aigle,M. *et al.* (2000) Genomic exploration of the hemiascomycetous yeasts: 3. Methods and strategies used for sequence analysis and annotation. *FEBS Lett.*, **487**, 17–30.
9. van Dongen,S. (2000) Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht.
10. Enright,A.J., van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
11. Llorente,B., Malpertuy,A., Neuveglise,C., de Montigny,J., Aigle,M., Artiguenave,F., Blandin,G., Bolotin-Fukuhara,M., Bon,E., Brottier,P. *et al.* (2000) Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. *FEBS Lett.*, **487**, 101–112.
12. Vandepoele,K., Saeys,Y., Simillion,C., Raes,J. and Van De Peer,Y. (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res.*, **12**, 1792–1801.
13. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
14. Stein,L.D. (2002) The Generic Genome Browser: A building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.